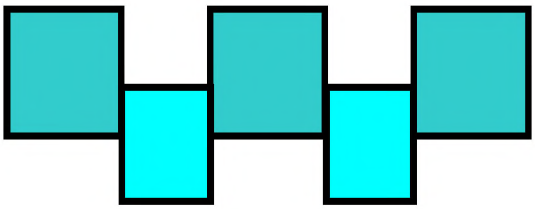
Authors: Louis Vuurpijl, Louis ten Bosch, Jan Peter de Ruiter,
Stéphane Rossignol, Lou Boves, Ralf Engel, Norbert Pfleger

Date: September 2002

Deliverable D4.1 Overview of the state of the art in fusion of speech and pen input



Document History

Version	Editor	Date	Explanation	Status
0.1	Engel	August 2002	Draft version to be read by the partners	Draft
0.2	Engel	September 2002	Incorporated comments from other partners.	Revisions
1.0	Engel	September 2002	PCC approval	Final

COMIC

Information sheet issued with Deliverable D4.1

Title:	Overview of the state of the art in fusion of speech and pen input
Abstract:	This document gives an overview over the current state of the art, including an overview over existing systems in the following areas: automated speech recognition (ASR), Automated gesture recognition (AGR), automated handwriting recognition (AHR), natural language understanding (NLU) and multimodal fusion.
Author(s):	Louis Vuurpijl, Louis ten Bosch, Jan Peter de Ruiter, Stéphane Rossignol, Lou Boves, Ralf Engel, Norbert Pfleger
Reviewers:	
Project:	COMIC
Project number:	IST- 2001-32311
Date:	September 2002
	For Public Use
Key Words:	automated speech recognition, automated gesture recognition, automated handwriting recognition, natural language understanding, multimodal fusion, multimodal dialogue systems, spoken dialogue systems
Distribution List:	
COMIC partners	Max Planck Institute for Psycholinguistics, Deutsches Forszungszentrum für Künstliche Intelligenz GmbH (DFKI), Max Planck Institute for Biological Cybernetics, University of Sheffield, University of Edinburgh, ViSoft, Katholieke Universiteit Nijmegen
External COMIC	IST, Public

Contents

Summary

This document gives an overview over the current state of the art, including an overview over existing systems in the following areas:

- ASR: Automated speech recognition
- AGR: Automated gesture recognition
- AHR: Automated handwriting recognition
- NLU: Natural language understanding
- Modality fusion

1 State of the art in Multi-modal Speech, Handwriting and Gesture Recognition: the COMIC perspective

Louis Vuurpijl, Louis ten Bosch, Jan Peter de Ruiter, Stéphane Rossignol, Lou Boves

This chapter describes the state-of-the-art in automatic speech, handwriting and gesture recognition (ASR, AHR, AGR) within the frame work of the target environments set out for COMIC. It is not intended to present a comprehensive overview over the achievements that have been made in multi-modal A?R throughout the years. It rather describes the major directions and approaches that have proven to be successful or that are currently being pursued. In particular, this chapter is concentrated on the current status of A?R in COMIC with the goal to provide the partners insight in the requirements, capabilities and restrictions that current A?R decoders impose on a multi-modal setting. The envisaged speech and handwriting/gesture recognition systems to be used in COMIC will be briefly described here. For an elaborate overview over the state of the art in multi-modal speech and pen input, we refer to the many excellent reviews in the literature, such as [Benoit et al., 1998,Oviat et al., 2001]. Furthermore, an annotated web site containing links to online documents about speech and gesture recognition, techniques used, research groups and commercial institutes involved, and successful systems and applications has been set up [Vuurpijl, 2002].

1.1 State of the art in automatic speech recognition

Automatic Speech Recognition (ASR) has become a broad field of research with a large number of specialized sub-disciplines. Worldwide, a substantial number of large groups work on ASR, both in academic and in company research environments. Universities such as CMU were among the first important players in the field. Nowadays, companies like IBM, SpeechWorks, Nuance, Microsoft, Philips, ScanSoft (L&H plus Dragon), 20/20 Speech, SpeechMachines, Vocalis, Softsound, VoiceSignal have taken the lead in industrially oriented commercial software for ASR. ASR results are presented at various conferences such as the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Eurospeech, and the International Conference on Speech and Language Processing (ICSLP). In addition, a large number of workshops are organized on various topics throughout the year.

Over the last years, the performance of Automatic Speech Recognition has shown a large improvement. The current ASR systems are performing well enough to be used in a wide variety of applications, such as interactive voice response, database queries over telephone lines, dictation, command and control, hands-free applications for telephones and navigation support in cars, voice-controlled internet access, interactive toys, etc. ASR is also a powerful tool to perform research in speech, signal analysis, phonetics, and dialogue modeling, and serves as a research tool for studies in performance robustness in adverse conditions, phonetic transcriptions, pronunciation variation, and dialect research.

An ASR system is built on three components: a dictionary enriched with phonetic transcriptions in terms of `speech models', acoustic models that relate acoustic sounds and these speech models, and a language model which tells the recognizer how to combine words into well-formed sequences.

Despite the progress in ASR performance, recognition results are far from 100 percent correct. Indeed, it has been estimated that the proportion of errors committed by ASR devices is almost an order of magnitude higher that recognition errors committed by

native speakers of a language (Lipmann, 1997). Typical performance results (word accuracy) that have been reported in the literature, in noise-free conditions, in single-speaker mode, broadband channel, are as follows:

Digits (10-12 words in the lexicon):	> 99 %
Command and control (100-200 words):	> 95 %
Free dictation (50000 and more words in the lexicon):	87-92 %

In adverse conditions, i.e. in noisy environments or over poor transmission lines, the performance may drop drastically, to 40-60% word accuracy for recognition of spontaneous speech over telephone lines. Broadly speaking, in noisy conditions humans perform a typical 10-15 dB better than machines do, across a large variety of experimental tasks (Lipmann, 1997). In other words: if an ASR device shows a word error rate (WER) of 5% in a specific task at a signal-to-noise ratio (SNR) of 15 dB, then humans will show 5% WER at a SNR of 0 dB. The performance level of an ASR system is determined by many factors. Most of these factors are related to a mismatch between training and test conditions, and are therefore relevant within COMIC as well. Important factors are the presence and type of environmental and channel noise, number of different speakers in the test, unexpected user responses, deviant speaking styles (shouting, whispering, fast speech), and the use of dialects or of particular accents.
ASR is often employed in multi-modal computer interfaces that include integration of modalities such as gestures, gaze tracking, lip reading, language understanding, handwriting recognition (Cohen et al., 1997, Oviatt et al., 1997, Vo et al. 1996, Marsic et al., 2000, see also the Proceedings of the ESCA Workshop on Interactive Dialogue in Multi-Modal Systems, 1999). For example, the SRI Multimedia Interfaces Group is building a series of prototype map-based applications that accept handwritten, verbal and gestural requests (see e.g. Cheyer et al., 1998).
Multi-modality is used to disambiguate user-generated information and/or simplify interaction. Both gestures and speech modality have their strengths and weaknesses. Gestures are convenient to draw a rough pictorial sketch or to graphically refer to objects. The speech modality is helpful to accomplish tasks that can be expressed in a formal way, such as `rotate this exactly 87 degrees counter-clockwise', and in the case when objects are invisible or unavailable.

1.1.1 Audio input: microphone

The microphone position is an important element in the design of the application, and needs to be carefully investigated. The main options are:

- Single microphone
    - As part of a head set, or attached to clothing. A head-mounted microphone is in general a good choice for noisy environments.
    - Fixed, at 30-60 cm from the speakers head, e.g. mounted on a table. A directional microphone of this type would be the best choice in noise-free `office' environments.
    - Fixed and attached to the application device. The problems here are background noise for internal drives and the usually poor quality of the built-in microphone.
    - Fixed and on a fair distance (1 m or further away). In this case, the signal processing has to take into account the echo characteristics of the recording room, and the background noise from behind the speaker.
- Microphone array. Arrays of microphones have a much higher spatial directivity than single microphones of the same type.

1.1.2 Noise

The presence of environmental background noise is detrimental for the recognition performance of an ASR system, if this noise has not been observed during training of the acoustic models. Also the type of noise is relevant: stable noise is less detrimental than e.g. babbling noise or impulsive noise and other non-stationary noises. Several methods are available to treat noisy conditions. One option is to train the acoustic models including the noise (this is probably the best option if the short-term noise spectrum is stationary and is the same in training and test). A second option is to attempt to reduce the noise component in the feature extraction process. A number of algorithms have been developed that attempt to clean the input acoustic signal by first estimating the noise spectral characteristics and next enhance the SNR and/or reduce the impact of the noise by e.g. spectral subtraction or spectral mean subtraction. The third option is confidence-based likelihood evaluation, in which signal/noise separation is dealt with in the model evaluation step in the ASR itself, by using confidence levels or a smart way of estimating the reliable parts in the signal. The University of Nijmegen has a large experience with noise reduction techniques.
In COMIC, we will aim at two options for noise reduction: the noise reduction in the feature extraction, and the confidence-based likelihood evaluation of the frame-state alignment score.

1.1.3 Multi-speaker tests

Multi-speaker performance is always worse than a system that is able to tune towards the characteristics of one speaker (by e.g. single speaker training - single speaker test). Two persons may differ in anatomical size and physical properties of the vocal tract and vocal folds, leading to possibly systematically different spectral characteristics for the same speech sounds. A speaker independent ASR typically shows a relative increase of word error rates of 150-200 percent compared to the single-speaker (speaker-dependent) case. This amount is more or less independent of the task.

1.1.4 Unexpected response to system prompts

Mixed-initiative or user-initiative dialogue management systems must allow the user to use a wider lexicon than system-driven dialogues. Often, a speaker does not know how to respond (especially infrequent commands), and a developer of the application cannot exactly predict how an average user will behave. Mismatches between the user's utterance and the language model in the system usually lead to a serious degradation of the ASR performance.
In fact, the quality of the acoustic models and the quality of the language model are of equal importance. During a dialogue, the language model can be adjusted on the basis of expectations about the user's behaviour in the next step of the dialogue by a feedback mechanism from the dialogue manager back to the ASR. However, estimation of the language model is difficult for the type of interactive dialogues tasks that are relevant in COMIC. The vocabulary and syntax strongly depend on the details of the task that a user intends to complete, as well as on personal habits and preferences of the users. Moreover, the amount of `text' that can be obtained by transcribing (COMIC-)dialogues is almost negligible compared to the billions of words that are used to train the language models in dictation systems. Fortunately, utterances produced in interactive dialogues tend to be relatively short and syntactically simple. That enables us to construct useful language models for dialogues in limited domains.

1.1.5 Accents and dialects

Accents and especially dialects usually decrease the performance of a speech recognition system that is trained on standard pronunciations (e.g. Schiel et al., 1998; Baum et al., 2001). In the last years, a number of techniques have been developed to adapt the phonetic transcriptions in the ASR lexicon by data-based adaptation methods. These methods for lexical adaptation work off-line. There is no stable, reliable method to adapt the phonetic representations on-line.

1.1.6 Conversational speech

ASR systems are usually trained on carefully pronounced speech, such as read speech or speech recorded from (semi-)professional speakers. This does not only relate to acoustic models, but also to lexical representations (i.e., the phonemic transcriptions of the words in the system's lexicon). It appears that the lexical representations (and probably also the acoustic models) developed for relative formal speech do not hold very well for conversational speech (the speech style that is customary in informal - and even in formal - human-human interaction). However, so far it has not been possible to develop reliable models of the differences between the lexical representations in `formal' and `conversational' speech styles. Consequently, recognition performance for extemporaneous and conversational speech is usually several orders of magnitude lower than for carefully pronounced `formal' speech.

1.1.7 Dialogue modelling

In systems requiring voice input such as a man-machine dialogue, the design of the overall system allows the state of the dialogue (as specified by the Dialogue Manager) to adapt the ASR towards the expected set of utterances in the next step. This feedback greatly enhances the performance of the ASR at that particular point in the dialogue. This improvement can e.g. be based on a reduction of the lexicon. For example, if the answer to a certain question must be a colour name, the ASR lexicon can be restricted to all colour names available from a background lexicon. Also an on-line modification (weighting) of the language model is feasible. There is, however, a trade-off between the performance of the ASR and the required flexibility in the responses by the user.

1.1.8 User satisfaction

The design of methods for performance evaluation is a major open research issue. The construction of predictive models of spoken dialogue performance, based on the combination of experimental data from different spoken dialogue systems, has been attempted. Combining several factors reported in the literature (Marsic et al., 2000; Larsen, 1999; Walker et al., 2000) we arrive at the following list:

- Execution time and response latency: the time it takes for the system to respond to the user input. In COMIC the delays of the modules will probably increase the total system delay. To avoid substantial response latencies at ASR side, we will investigate the possibility of early output of provisional hypotheses.

- Time spent in subtasks. This is related to execution time but measured over a meaningful span of commands.

- Number of turns within subtasks.

- The correctness of the execution of the command. This relates to the number of speech recognition errors by the ASR and the corrective power of subsequent meaning extraction module and the dialogue manager. Shortcomings or ambiguities of the ASR output can be resolved or reduced by various techniques. (a) a proper dialogue handling can provide context information and an expectation of the set of words a user is about to use in the next dialogue step. (b) the ASR output can be processed by using additional knowledge sources, such as a detailed domain description or an elaborate syntax. A dialogue system can receive positive scores despite performance problems with the speech technology used in the implementation (Boves & den Os, 1999)

- The successful completion of the dialogue

- The `ease of use' as experienced by the (naive) user

These factors are known to be important in the overall performance ranking of dialogue systems. For ASR, these factors imply that correctness as well as response time are

crucial. Therefore, feedback information on context from the dialogue manager is relevant to enhance the performance of ASR at specific points in the dialogue.

1.1.9 ASR implementation in COMIC

In COMIC, an ASR module will serve as speech-to-text system at the input side, translating the acoustic information of what has been uttered into a word lattice (enriched with confidence levels, and with prosodic information). The format of the output file will be XML-like.
A feasible option for the ASR `engine' is the HTK source base. HTK is open source, free of charge. Research data obtained by using HTK can be re-used freely in scientific papers and commercial products. HTK is supported worldwide, and there is an active research community using the software. Although originally devised for UNIX/Linux, there is increasing interest (also shown by remarks in the HTK users group on the internet) to have HTK running on a variety of Windows platforms.
At the KUN, HTK has recently been ported to Windows by Arjan Lamers to a i386 architecture. The executables have been compiled by using the minGW environment, after slight adaptations of the unix make-files. The ported HTK version is v3.0. The latest stable version that is available now (March 2002) via the HTKwebsite is v3.1.
From version 3.0 on, HTK has vocal track length (VTL) normalisation including full variance transformation. Duration parameters are not used during decoding. The module VInit performs a Viterbi training (which is faster than and probably preferable over a time-consuming Baum-Welch training); and HRest performs a Baum-Welch training (which is more time-consuming).
For the COMIC application a number of issues must be addressed. Below an overview is presented.

Speech input, microphone

In HTK, speech can be read from a file or obtained via direct audio. Using file input, there is no buffering. Direct audio uses a buffer that is now 3 seconds long. Feature extraction processing runs a few frames behind - but conceptually in parallel with - speech input.
DFKI uses various microphones for SmartKom demonstrations (e.g. a Sennheiser ew 152 headset). Earlier recordings have been made with a table-mounted microphone. The SLOT hardware allows to use one or two clip-on microphones attached to clothing. The ViSoft application will make use of a high-quality close-talk microphone.

Flexibility of the direct audio mode

HTK has a recognition mode using direct audio input. In that case, there is no energy normalisation. For training models for live audio input, energy normalisation should be off. The input data sampling is governed by a configuration parameters SOURCERATE, which can also be set by an audio control panel. In the latter case, it must be possible for HAudio (the HTK module that takes in audio input) to obtain the sample rate from the audio driver. The HAudio module provides two facilities for audio input control. (A) The first method involves an automatic energy-based speech/silence detection, which is enabled by a specific parameter setting in the configuration file. (B) The second mechanism for controlling audio is by arranging for a signal to be sent from another process. The signal toggles the activation of the audio input device. (A) and (B) can be combined such that a signal activates audio input while the recording is stopped after a silence detection or after a second signal. HAudio also has a replay buffer.

Early hypotheses output

The actual search in HTK is performed by a module HRec using a token passing mechanism. The basic usage just involved single best and back traces on word level. In COMIC, the use of N-best lists and/or word lattices will probably be essential. This is well possible in HRec. Whether provisional recognition results can already be presented during the processing of an utterance will be investigated during the COMIC project. The

word lattice is constructed on the basis of information in the heap, which is incrementally updated over time, and is reset after the (previous) hypotheses have been outputted.

Incorporation of prosodic information

In the mainstream approaches to automatic speech recognition, prosody does not play a prominent role. Despite a long tradition of experimentation with the use of prosodic information in speech recognition, these experiments have been largely unsuccessful from the point of view of improving transcription accuracy. However, during the recent years it has been attempted to incorporate aspects of prosody, such as pitch and 'prominence', into the ASR paradigm in order to provide the recogniser with more information and to obtain improved recognition results. Successful effort has been undertaken to find ways of incorporating prosodic information into a wider variety of ASR-related tasks, such as identifying speech acts, locating 'important' words and phrases, improving rejection accuracy, finding topic segment boundaries, locating and processing disfluencies, identifying user corrections in dialogue, identifying speakers and languages, and detecting speaker emotions. For all of these research areas, the collection of appropriate corpora and the development of useful prosodic labelling schemes for them have been critical. In the ASR feature extraction, prosodic features such as pitch, speaking rate, phrasing, emphasis are not used. However, during the recent years it has been attempted to incorporate aspects of prosody, such as pitch and 'prominence', into the ASR paradigm in order to provide the recognizer with more information and to obtain improved recognition results. Prosody can be helpful in a number of ways. Prosodic information might disambiguate utterances and influence their function in a dialogue. When produced in a dialogue, the utterance 'No, I'll take the train to London at 5 PM' can have substantially different meanings depending on the prosodic realisation (especially intonation). In general, prosody can be an essential auxiliary factor in NLU (natural language understanding), confirmation and clarification, and a factor of importance to put emphasis on specific information. Furthermore, prosodic information might help is to distinguish words with a similar pronunciation (such as 'record, re'cord), and to identify focus.
With respect to the incorporation of prosodic information, the simplest thing will probably be to write, in parallel with the regular feature extraction (FE), a specialized dedicated FE to extract pitch and energy information, and to use this second stream for getting cues about prosodic topic marking. It is probably not a good idea to include the prosody directly into the speech recognizer, since the prosodic information that we are aiming at is mostly suprasegmental, or even far beyond segmental level. The usefulness of the energy component depends on the specific normalisation techniques on utterance level, the position of the microphone, and the position changes of the speaker. How to measure the amount of hesitation is to be studied (usually hesitation manifests itself in a combination of phonetic factors, phrasing behaviour, and lexical issues, and so is `multi-modal' by itself).

Databases for ASR training and test

The demo version for VIsoft aims at two target languages: English (British English) and German. Since COMIC defines two different platforms, one `research' platform and one fully integrated non-commercial demonstration platform, one may envisage different ASR settings for training and test for optimal performance. Since the VIsoft application is likely to be demonstrated in realistic (noisy) environments, commonly applied adaptation techniques will be used in the recognition phase, such as noise reduction and on-line channel adaptation.

Multimodality v.s. unimodality

The combination of the three different modalities speech, gestures, and handwriting (ASR and AGR and AHR) might increase the ease of use of an application, and may form an auxiliary factor to reduce the ambiguity in the interpretation of the individual input channels. A multimodal system has in one way or another to take care for the fusion component, in which the fusion of information can take place at the level of data, features

or decisions (Marsic et al., 2000). Systems based on late fusion combine on the decision level. This is one of the options in COMIC.

1.1.10 Within the near future

One of the most important factors in the progress of ASR performance during the last 20 years is the progress made in the processing capabilities of CPUs and memory chips. The increased processing power enabled the use of complex, high CPU- and memory intensive calculation schemes. Also the growth in the size of speech and text corpora substantially contributed to the better modelling of acoustic and language models. The performance gain in ASR is to a smaller extent based on the increase in fundamental knowledge about speech or the intrinsic improvements of the algorithms used in training and word search. Actually, the techniques that are currently used in mainstream ASR approaches seem to be quite crystallized, although a number of research groups continue to search for new approaches: alternative feature extractions, search methods, alternative training criteria, and new combination with other knowledge bases. A major progress is to be expected in the field of dialogue modelling and multi-modal interaction, since many details are still to be unraveled about human multi-modal behaviour while these details are important for a proper understanding and modelling of machine-mediated human-human interaction or man-machine interaction.
The improvement of ASR performance in the near future is liely based on a larger CPU capacity at the customer's side (providing opportunities for a smart preprocessing front end that were out-of-reach only a few years ago), on better systems for noise reduction, on new adaptation techniques, and on an adequate integration of semantically related information. Also the improved combination of ASR and additional knowledge sources will be an important factor.
In the context of dialogue systems such as COMIC, the most important factors that enhance the usability of ASR is the improvement of the response time of the ASR module, and the ability of the system to feedback to the ASR the expected semantic context of the next utterance. The response time can be shortened by generating early meaningful hypotheses after processing a part of the input acoustic signal. A proper feedback from the dialogue manager to the ASR will enable the ASR to constrain the test lexicon in a proper way to minimize the number of word errors during recognition.

1.1.11 References and selected Bibliography for ASR

Speech technology in general

IEEE Special issue on Language Processing (2000). Proceedings of the IEEE, vol. 88, no. 8.

Jurafsky, D. & Martin, J.H. (2002). Spech and Language Processing. An introduction to natural language processing, computational linguistics and speech recognition. Prentice Hall, New Jersey.

Lippmann, R.P. (1997). Speech recognition by machines and humans, Speech Communication, 22(1), 1-16.

Usability (selection)

Boves, L. & den Os, E. (1999) Application of speech technology: designing for usability. *Proceedings of the ASRU-1999 conference*, Keystone.

Larsen, L.B. (1999). Combining objective and subjective data in evaluation of spoken dialgues.Proceedings of the ESCA Workshop on Interactive Dialogue in Multi-model Systems, Kloster Irsee, Germany, 89-92.

Marsic, I.., Medl, A. & Flanagan, J.L. (2000). Natural Communication with Information

Systems. *IEEE Special issue on Language Processing. Proceedings of the IEEE*, vol. 88, no. 8. 1354-1366.

Walker, M. A., Kamm, C. A., & Litman, D. J. (2000). Towards developing general models of usability with PARADISE. Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems.

Noise robustness, KUN papers (selection)

de Veth, J., Cranen, B., Boves, L., 1998. Acoustic backing-off in the local distance computation for robust automatic speech recognition. In: Proc. International Conference on Spoken Language Processing, Vol. 4, pp. 1427-1430.

de Veth, J., de Wet, F., Cranen, B., Boves, L., 1999. Missing Feature Theory in ASR: Make sure you miss the right type of features. Proc. Workshop on Robust Methods for ASR in Adverse Conditions, pp. 231-234.

de Veth, J., Cranen, B., de Wet, F., Boves, L., 1999. Acoustic pre-processing for optimal effectivity of missing feature theory. Proc. Eurospeech-99., pp. 65-68.

de Wet, F., Cranen, B., de Veth, J., Boves, L., 2000. Comparing acoustic features for robust ASR in fixed and cellular network applications. In: Proc. ICASSP-2000, pp. 1415-1418.

de Veth, J., Cranen, B., Boves, L., 2001. Acoustic features and distance measure to reduce the vulnerability of ASR performance due to the presence of a communication channel and/or background noise. In: 'Robustness in Language and Speech Technology'. J.-C. Junqua & G. van Noord (Eds.), Kluwer, Dordrecht, 9-45.

de Veth, J., de Wet, F., Cranen, B., Boves, L., 2001.Acoustic features and distance measures that reduce the impact of training-test mismatch in ASR. Speech Communication, 34 (1-2), 57-74.

de Veth, J., 2001 On Speech Sound Model Accuracy. PhD Thesis, University of Nijmegen, April 10th, 2001.

de Veth, J., Cranen, B., Boves, L., 2001. Acoustic backing-off as an implementation of Missing Feature Theory. Speech Communication, 34 (3), 247-265.

de Veth, J., Mauuary, L., Noe, B., de Wet, F., Sienel, J., Boves, L., Jouvet, D., 2001. Feature vector selection to improve ASR robustness in noisy conditions. In: Proceedings Eurospeech 2001, pp. 201-204.

de Wet, F., Cranen, B., de Veth, J., Boves, L., 2001. A comparison of LPC- and FFT-based acoustic features for noise robust ASR. In: Proc. Eurospeech 2001, pp. 865-868.

van de Werff., L., de Veth, J., Cranen, B., Boves, L., 2001. Analysis of disturbed acoustic features in terms of emission cost. In: Proc. Workshop on Consistent & Reliable Acoustic Cues for sound analysis (CRAC), 4 pages, no page numbers.

Pronunciation Modeling (selection)

M. Baum, R. Muhr & G. Kubin (2001). A Phonetic Lexicon for Adaptation in ASR for Austrian German. In Proceedings of ISCA Workshop "Adaptation Methods for Speech Recognition", Sophia-Antipolis.

F. Schiel, A. Kipp, H.G. Tillmann (1998). Statistical modelling of pronunciation: it's not the model, it's the data In: H. Strik, J.M. Kessens, M. Wester (eds.), Proc. of the ESCA workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition', Rolduc, Kerkrade, 4-6 May 1998, pp. 131-136.

Pronuciation Modeling, KUN papers (selection)

Strik, H. & Cucchiarini, C. (1999). Modeling pronunciation variation for ASR: a survey of the literature, In: Strik et al. (eds.), Speech Communication 29, 225-246.

M. Wester (2002). Pronunciation Variation Modeling for Dutch Automatic Speech Recognition, Ph.D. thesis, University of Nijmegen, The Netherlands.

J.M. Kessens & H. Strik (2001). Lower WERs do not guarantee better transcriptions. Proceedings of Eurospeech 2001, Aalborg, Denmark, vol. 3, pp. 1721-1724

M. Wester, J.M. Kessens & H. Strik (2000)Using Dutch phonological rules to model pronunciation variation in ASR PHONUS5: Proceedings of the Workshop on Phonetics and Phonology in ASR. Saarbuecken: Institute of Phonetics, University of the Saarland, pp. 105-116

M. Wester & E. Fosler-Lussier (2000). A comparison of data-derived and knowlegde-based modeling of pronunciation variation. Proceedings of ICSLP 2000 (Volume I), Beijing, China, pp. 270-273.

Confidence measures, KUN papers (selection)

A.G.G. Bouwman, J. Sturm & L. Boves (2001). Effects of OOV rates on Keyphrase Rejection Schemes. Proceedings Eurospeech 2001, Aalborg, Denmark, 2001, pp. 2585-2588

A.G.G. Bouwman & L. Boves (2001). Using Discriminative principles for recognising City Names Proc. ITRW on Adaptation Methods for Speech Recognition, Sophia-Antipolis, France, 2001, pp. 109-112.

A.G.G. Bouwman & L. Boves (2001). Using Information on Lexical Stress for Utterance Verification. Proceedings of ITRW on Prosody in ASRU, Red Bank, 2001, pp. 29-34.

A.G.G. Bouwman, L. Boves & J. Koolwaaij (2000). Weighting Phone Confidence Measures for Automatic Speech Recognition. COST249 Workshop on Voice Operated Telecom Services, Ghent, Belgium, pp. 59-62.

A.G.G. Bouwman, J. Sturm & L. Boves (1999). Incorporating Confidence Measures in the Dutch Train Timetable Information System Developed in the Arise Project. Proceedings. International Conference on Acoustics, Speech, and Signal Processing, Phoenix, USA, vol. 1, pp. 493-496.

Multimodality

Marsic, I. & Attila, M. (2000). Natural communication with information systems. Proceedings of the IEEE, vol. 88 (8), 1354-1366.

Oviatt S., DeAngeli, A., & Kuhn, K. (1997). Integration and synchronization of input modes during multimodal human-computer interaction. Proceedings Conf. Human Factors in Computing Systems (CHI'97), Atlanta, 415-422.

Proceedings of the ESCA Workshop on Interactive Dialogue in Multi-Modal Systems, 1999.

Prosody (selection)

Swerts M, Terken J. (eds). (2002). Dialogue and prosody. Special double issue of Speech communication on dialogue and prosody, 36:1/2.

Krahmer E., Swerts M, Theune M., Weegels M. (2002). The dual of denial: two uses of

disconfirmations in dialogue and their prosodic correlates. Speech communication, 36, 133-145. ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding (2001) The Molly Pitcher Inn, Red Bank NJ, October 22-24, 2001.

1.2 State of the art in pen input recognition

In COMIC, pen input is concerned with the recognition of a trace of subsequent coordinates which are captured from a range of input devices, such as touch screen, (LCD) tablets or mouse. Pen input recognition is a technology for the recognition of handwritten words, characters, digits, symbols, or gestures. The recognition of the first four classes can be labeled as automatic handwriting recognition (AHR) and the last class as automatic gesture recognition (AGR). An important research topic to be considered in COMIC is to study ``mode management''. In a context where the application intermixes (and supports) AHR and AGR, it has to be determined which mode is intended by the user. For that reason, pen input recognition will have to be able to handle and distinguish multiple modes, a topic which has not been extensively studied until now.
AHR is a field of research that is concerned with the recognition of handwriting. Although AHR is practiced by researchers from various disciplines, results are mainly presented at two conferences: the ICDAR (international conference on document analysis and recognition) and IWFHR (international workshop on frontiers of handwriting recognition). The latter series of conferences is particularly focused on so-called *online* handwriting recognition, i.e. technologies for recognizing the dynamic trajectory that writers produce with their pen tips and that is acquired through digitizers, or writing tablets. The ICDAR conferences are dominated by contributions from *offline* handwriting recognition, although they also contain papers concerned with online AHR. In offline AHR, handwriting is scanned from paper and image processing and pattern recognition techniques are used to interpret the optically scanned handwriting. A small amount of papers presented in this community is concerned with gesture recognition (AGR). AGR is studied from the perspectives of pattern recognition and (multi-modal) human-computer interaction. There are several conferences (partly) devoted to AGR such as the international conference on automatic face and gesture recognition or the gesture workshop. References to both AHR and AGR can be found in journals concerned with pattern recognition, computer graphics, computer vision and man-machine interaction. In COMIC, we study interactive man-machine communications and therefore, only *online* handwriting and gesture recognition are examined.
A large number of research labs is concerned with AHR and numerous interactive applications have been built that exploit AHR since the last two decades. It is beyond the scope of this report to provide an overview over all commercial AHR systems available today. Furthermore, commercial firms do not report recognition rates, which makes a comparison on handwriting recognition systems very difficult. In the latest issue (July 2002) of Pen Computing, the dominant magazine that reports on the latest developments in PDA's and tablet computers, users and reviewers of handwriting recognition software seem to agree on the subjective opinion that the Microsoft recognizer that comes with Windows XP for tablets is *the best they've ever seen*. Based on my personal experience with PDA's and the system from Microsoft, I would share this opinion (LV). However, it is clear that the problem of writer-independent handwriting recognition is not solved yet, as all systems only reach an acceptable performance if the owner takes the effort of training them with his own handwriting. Commercial systems using AGR are not widely spread.
In general in AHR, the set of classes that can be recognized is known in advance, e.g. the characters from the alphabet or the words occurring in the lexicon. This is in contrast to AGR, where the gesture classes are not so well-defined. For handwritten sentence recognition, techniques well-known from the speech community are used for recognition, such as the use of language models or statistical information about the occurrence of words. In COMIC, the determination of a set of meaningful gesture classes in the context of route map scenarios and bathroom decorations is a research issue. Here, we would like to make a distinction between 2D gestures (captured by a digitizing tablet), 2.5D gestures (captured in a limited space above the tablet, e.g. with the pen-tip in the air) and 3D gestures (captured with, e.g., optotrak, glove or video equipment).
The examination of 2.5D and 3D gesture recognition will be pursued in a later stage of the project. In COMIC, knowledge about the physical constraints that restrict human

movements may be used as hints for 3D gesture recognition. Much is known about the human handwriting production process and new 3D tracking techniques are now being used to validate models of human 3D motor control. The dScript AHR recognizer, which is briefly described in this report below, builds on assumptions about the handwriting production process, e.g. that handwriting is constituted from ballistic pen movements and that in handwriting, humans are explicitly taught how to acquire legible and efficient handwriting skills. A number of excellent groups in the world have worked on this topic, in particular research teams lead by Plamondon, Thomassen, and Morasso. A recent PhD study by Breteler [Klein Breteler, 2001] examines biophysical and cognitive determinants of human trajectory information. The outcomes of her thesis provide hope for 3D gesture recognition, as she concludes (based on experimental optotrak data) that though the degrees of freedom to produce 3D movements are high, humans seem to have a limited number of preferred trajectories. Please note that if it were possible to extrapolate an intended trajectory based on such models and a number of early registrations from the trajectory, it may even be feasible to consider the output of early gesture classification hypothesis *before the gesture is terminated*.

1.2.1 Progress in online AHR

Automatic handwriting recognition operates on ``digital ink'', i.e. a still image captured through some scanning device (offline AHR), or a dynamic trajectory comprising a trail of (x,y) coordinates produced by a pen+digitizer, mouse or touch screen (online AHR). In general, online AHR has the advantage that the dynamics of the pen trail is available and thus yields a higher recognition performance than OCR systems. For multi-modal interactive systems, online handwriting recognition is required. Therefore, we will discuss online AHR in this chapter, though it will be indicated that certain techniques borrowed from offline AHR may be of use for COMIC. For a review on offline AHR, we refer to, e.g., [Steinherz et al., 2002,Vinciarelli, 2002]. A recent review on online handwriting recognition is presented in [Plamondon et al., 1999].
Similar to ASR, the performance of an AHR system depends on a number of factors, such as lexicon size, writing style, input device and sampling rate, though AHR is less sensitive to environmental noise. As is shown in [Srihari et al., 2001,Vuurpijl et al., 2002a], the between-writer variability in handwriting is high, resulting in the effect that each new writer will introduce new shapes. This problem makes the comparison of AHR systems even more difficult. The latest ICDAR and IWFHR conferences, held in Seattle, Amsterdam and Niagara on the Lake, have set the current state of the art. Some authors claim recognition rates of more than 98% for isolated character recognition. The performance of digit recognizers is even higher. However, isolated word recognition in a writer-independent fashion yields results that may fall back below 75% in sloppy, cursive handwriting conditions. One of the best online writer-independent AHR systems available [Jaeger et al., 2000] yields recognition rates of 96%, 93.4% and 91.2% for respectively a 5K, 20K and 50K lexicon. This advanced system called NPEN++ was trained on three databases comprising in total 488 writers producing more than 28000 words, and builds on the convolutional neural network and HMM approach described in [Bengio et al., 1994].
The dScript system developed at the NICI is trained with the UNIPEN database, containing handwriting information of more than 2200 writers. For a 20K word lexicon, recognition rates of 85% are achieved, where it should be noted that the UNIPEN database is considered as particularly difficult data. UNIPEN has been the de facto standard in online handwriting data formats since 1994 [Guyon et al., 1994]. More than 40 institutes contributed to the data collection, which involves that it comprises data collected on different tablets (resolution, sampling rate), writers and writing conditions. From the same database, dScript achieves more than 96% accuracy on classifying isolated characters, which until recently was considered as the state-of-the-art. More recently, recognition results of more than 97% are reported on the same data [Parizeau et al., 2001]. The highest recognition rates for digit recognition on the UNIPEN dataset are reported in [Ratzlaff, 2001], with a score of 98.8%.

1.2.2 Pen input devices and operating systems

A wide range of input devices are available to capture digital ink. Using devices such as the light-pen, LCD-tablet or touch screen, the user has direct control with his stylus or finger by pointing to visible objects on the screen. If the user uses the mouse, pressure tablets or digitizing tablets, this requires a certain amount of hand-eye coordination as there is no immediate visual feedback. Note that this is more intrusive than the natural interactions that are pursued in COMIC. Other systems using data gloves or optical data acquisition tools such as (infrared) video cameras (finger tracking) or optotrak require even more efforts from the user to interact with the system. In [Benoit et al., 1998], the following technical requirements are listed for deciding on a pen input device: visual feedback on handwriting movement, sampling rates from 50Hz or more, spatial resolution of at least .1 mm and the possibility to distinguish x/y position, pressure and pen-up/pen-down information. Please note that the proposed input device for COMIC: the Wacom Cintiq 15x fulfills all these requirements. Future small mobile target platforms to be used in COMIC will probably not meet all requirements, but the data sheets for the current scanning technology produced by Fujitsu [Fujitsu, 2001] show that these technical requirements will most probably be met if a tablet computer such as the Fujitsu-Siemens Stylistic would be used.
In a recent issue of Pen Computing (July 2002), a review on tablet computers is presented, including the announcement of the tablet pc code-named *MIRA*, that was promoted at the IWFHR8 conference August 2002 by the handwriting recognition team from Microsoft. This new device runs Windows XP (tablet edition) and it is expected that the MIRA will significantly boost the market for pen-computing applications such as *inkMail* or *ink messaging*. Up until now, tablet computers and PDA's used a dedicated operating system, such as PalmOS or Windows CE. Such operating systems are tightly coupled to the hardware, and provide application interfaces that can be used by software developers to build applications, e.g. by using the captured online ink or by interfacing to the handwriting recognition engines that are shipped whith the hardware.
The MIRA is supported by the top management from Microsoft and will include pen-input for office applications such as PowerPoint. The MIRA is manufactured by a range of hardware firms and is expected to be released at the end of 2002. Specifications of the MIRA are due to be released, but given the expected support from several Microsoft product lines, it is expected that the tablet PC will certainly become a popular device that is a suitable target platform for the proposed COMIC applications. Therefore, it should be considered to use Windows XP (next to Linux) as the target operating system for COMIC. Moreover as Microsoft has ceased to support Windows 2000.

1.2.3 Techniques used in AHR

AHR is recognized as a difficult pattern recognition (PR) problem and as such performs the well-known stages in PR: data acquisition, preprocessing, normalization, segmentation, feature extraction and classification [Vuurpijl, 2002]. Preprocessing the trajectory of x,y coordinates is concerned with low-pass filtering (smoothing), interpolation (in cases where the input device misses samples) and spatio-temporal resampling of the data. Normalization operations involve the correction of size, slant and rotation. These operations are closely connected to the problem of baseline detection. Most AHR systems available today and in particular the NPEN++ and dScript systems use similar preprocessing and normalization techniques. As the detection of baseline, slant, corpus, ascender and descender heights is more reliable for words (or sentences), most recognizers only output a classification result after completion of a word.

1.2.4 Features in AHR

As offline features in general provide an orthogonal view to handwriting (as opposed to online features), today, both online and offline features are computed from the online handwriting signal. In particular for bar crossings (t,f,z,..), dots (i,j) and other examples where the stroke order in handwriting introduces confusion (like in the E), the still image may provide the required evidence for disambiguating between character classes. Both NPEN++ and dScript use offline features based on the bitmap generated from the online

signal. It is beyond the scope of this chapter to enumerate all possible features in AHR. However, as we will examine both AGR and AHR in COMIC, a number of suitable features for gestures and handwriting will be listed here. A general distinction can be made between local and global features. Global features are, e.g., ascenders, corpus letters and descenders (ascordesc pattern), width, height, number of dots, number of crossings, etc. Typically, these features are used for holistic template matching of words. Given the sequence of (x,y,p) coordinates, the local writing direction, curvature, velocity, angular velocity, slope, etc. are known features. Besides, the dScript system uses 14 stroke-based features as described in [Schomaker, 1993]. Although in principle, many of these local features are independent of the writing language (in fact, many of them may be used for other alphabets than western handwriting), some of the features are language dependent. In particular, this holds for curliness, slant and transitions between subsequent characters. The latter involves that handwriting systems based on local features will also require re-training if a new language is introduced.

1.2.5 Template-based character and word recognition

Given a number of prototypical word or character shapes with corresponding extracted features, a AHR system can match an unknown exemplar to the list of known shapes via template matching. Any classification technique like neural networks, k-means clustering, HMM or nearest neighbour methods can be exploited for this goal. Note that given a gesture repertoire with a proper set of extracted features, this same paradigm can be used for AGR. In COMIC, we will first pursue this idea for AGR, in the context of the SLOT experimentation paradigm and the ViSoft bathroom application. It is hoped that after transcription of the first data collections, a preliminary set of gesture prototypes will become available that is suitable for template matching. Template matching is constrained by the shape repertoire, so if a new shape class is added, this may involve that new features have to be invented/computed and/or classifiers have to be retrained. On the other hand, the technique makes it possible to explicitly describe prototypes in handwriting, as described in [Vuurpijl and Schomaker, 1997]. Furthermore, one of the geometric classifiers in dScript uses a grammar for describing certain character shapes (such as "L=<stroke down; stroke right>"), an approach that may be suitable for unconstrained gesture recognition as will be indicated below. And finally, a well-understood inventory of shapes in handwriting can be of use for the computer rendering of ink, which may be a topic of research in the frame work of COMIC.

1.2.6 Analytical approaches

Handwritten sentences are constructed from words, words from characters and characters are produced by a sequence of strokes. The analytical approach builds on this hierarchy of information levels by zooming in at the input, recognizing, e.g. strokes at the lowest hierarchical recognition level, subsequently recognizing characters, words and sentences. In [Benoit et al., 1998], analytical approaches are further distinguished based on the exploitation of explicit or implicit segmentation techniques. The first techniques segment and recognize characters from the input and subsequently match the resulting sequences of characters to the words in the lexicon. Implicit segmentation techniques match character and word-level information in a single step to the word models from the dictionary. Techniques well-known from speech recognition can be used to match sequences of words to sentences, making use of language models or statistical information about word sequences. If any technique in AHR can be used to output early probabilistic recognition hypotheses, it must be the analytical approach with explicit segmentation. Note that analytical approaches can very well exploit template matching techniques for matching input segments to known handwriting shapes. As it may be expected that handwritten words or sentences will not often be produced in COMIC, the effects of skipping traditional word normalization techniques will be examined.

1.2.7 Multiple classifier systems (MCS) and multi-modal integration

Contributions of specialists in AHR to the IWFHR and MCS series of conferences and also to the upcoming special issue of the IJDAR on multiple classifier systems show that

also in AHR, the concept of combining the expertise of multiple classifiers is a trend. MCS have proven their value in all areas of pattern recognition and also in AHR, reductions in error rate by a factor 2 or more can be achieved [Duin and Tax, 2000]. It has been debated whether techniques from MCS may be exploited for multi-modal classifier combination. In uni-modal systems, multiple classifiers operate on the same given input signal. In multi-modal systems, information has to be fused from heterogeneous sources. For example in [Wu et al., 1999], it is stated that only in specific situations, signals can be integrated at the feature level (i.e. in speech and lip movements). On the semantic level, the combination of signals makes only sense if they refer to the same multi-modal concept. An interesting paradigm for analyzing references from different modalities to the same concept is what they call the associative map. The paper of Wu et al is one of the few papers that statistically analyze multi-modal fusion.

1.2.8 The dScript recognition system at NICI

The NICI has been involved in handwriting research for almost two decades. The motor control group headed by prof. Thomassen and prof.van Galen has studied the production of handwriting. Professor Schomaker, now affiliated with the Groningen University, has led the (still ongoing) research on the recognition of human handwriting. At the latest IGS conference (international graphonomics society), it appeared that it is still valid to consider the VBS (velocity based stroke) as the basic building block in handwriting [Schomaker and Teulings, 1990]. Human handwriting is thus, a sequence of VBS and AHR can be solved by finding the sequence of stroke-transitions that best matches a given exemplar of handwriting. The AHR system developed at the NICI still uses the VBS. The approach is general purpose in the sense that it can be used to recognize any kind of handwriting, e.g., not only western but also Arabic, Hangul, Katagana or Sanskrit. During the last five years, the system has been equipped with a multiple of classifiers, such as neural networks, HMM, support vector machines [Vuurpijl and Schomaker, 2000,Wang et al., 2000] and clustering algorithms [Vuurpijl and Schomaker, 1997]. The output hypotheses from classifiers are merged through the concept of multiple agents [Vuurpijl and Schomaker, 1998] and several weighting schemes. A few years ago, the system was considered state-of-the-art, but as indicated above, today other systems achieve better recognition performances than dScript, although no real between-system benchmark tests were held on the same data.
For COMIC, a new system based on selected components from dScript will be developed. This new system will recognize online characters, digits, words and parts of it will also be used for the recognition of pen-based generated gestures. A research question to be pursued is the early determination of generated classes, i.e., based on what part of the handwriting input can it be decided whether a gesture is produced, a character, word, or graphics?
Also, the exploitation of pen-pressure for AHR has not been studied in detail yet. In COMIC, the variability in pen-pressure and velocity may be used as a clue for the determination of states of human stress or anxiousness.

1.2.9 The UNIPEN database of online handwriting and inkXML

Until 1999, NIST has been the organizing party in the collection of a large amount of handwriting data, called UNIPEN [Guyon et al., 1994]. As of 1999, the international unipen foundation was installed to safeguard the distribution of a large database of on-line handwritten samples, collected by a consortium of more than 40 companies and institutes. Over 5 million characters have been collected, from more than 2200 writers. The data can be obtained on CDROM via the IUF web site (IUF). In the mean time, UNIPEN has evolved as one of the de facto standards in online database formats. There are now various initiatives to develop new, XML-based, standards to specify online handwriting, provisionary called inkXML. The IUF is involved in these initiatives. For COMIC, the UNIPEN database is sufficient for training English word recognizers. The amount of German words in UNIPEN is very limited which involves that it may be required to acquire German databases for training German word recognizers.
At the IWFHR8 workshop (August 2002), an inkXML tutorial was presented by representatives from Motorola and IBM. A small group has been working on a draft

specification of inkXML, which will become part of the multi-modal interaction group of the W3C consortium. As such, inkXML will become a standard, next to voiceXML. Drafts of the specifications will be downloadable from the W3C websites at the end of August 2002. It appeared that the currently used Microsoft ink format does not differ much from the proposed formats. As Microsoft is involved in the MMI group of the W3C (voiceXML), it is not unthinkable that its handwriting group will consider inkXML.
Within the frame work of the research conducted at institutes involved in AHGR, various other online databases have been collected over the world. An overview over these databases and institutes concerned with AHGR is listed in [Vuurpijl, 2002]. In the frame work of COMIC, these groups will be contacted for sharing information and data.

1.3 State of the art in automatic gesture recognition

In COMIC, we will examine pen-based gesture recognition, distinguishing between inputs captured through a digitizing tablet (2D) and, in a later phase of the project, inputs captured through 3D acquisition devices. There is a clear distinction between the two inputs. The first can be considered as a special case of AHR. Given a set of gesture classes, template matching or analytical techniques can be used to perform recognition. Most gesture recognition systems still use the kernels developed by Rubine [Rubine, 1991] and it is stated in [Benoit et al., 1998] that a gesture recognition system based on template matching "can be implemented in a few days".

1.3.1 Towards natural 2D gesture recognition

The approach of teaching the user a number of gestures that can be recognized is also known as scribble matching [Fonseca and Jorge, 2000] and is often used in drawing tools [Julia and Faure, 1995] graphical editors, or handwriting applications where users have to learn a new alphabet comprising shapes that are relatively easy to recognize (cfr. the unistroke alphabet from Goldberg [Goldberg, 1989]). Please note that in COMIC we will study natural man-machine interactions, where it is intended that users are not constrained by a pre-defined set of gestures. The recognition of 2D shapes has also extensively been studied in the context of computer graphics and image retrieval [Veltkamp and Hagedoorn, 2000]. This research has resulted in new feature extraction techniques for object recognition and a theory called curvature scale space [Mokhtarian et al., 1996,Lindeberg, 1998], which are now being considered for the Vind(x) image retrieval system developed at the NICI [Schomaker et al., 1999,Vuurpijl et al., 2002b]. Vind(x) exploits a "new use for the pen", i.e. object recognition through outline sketch. Given this wealth of techniques for recognizing sketches, outlines or scribbles, it is not unthinkable that natural gesture recognition is a solvable problem indeed. For a limited amount of gesture classes (<10), recognition rates of 95% to 100% are very well feasible.
Although scribble or outline matching may be a useful approach if the user is willing to produce one out of a set of predefined shapes, in COMIC we will mainly consider natural gesture recognition. Gestures produced in a natural dialog between two humans or between a human and the machine, will most probably introduce more classes and a much higher within-class/between-subject variability than in the context of constrained gesture recognition. Although the work performed by Breteler [Klein Breteler, 2001] suggests otherwise. Please note that her study studies functional behavior (e.g., hitting a light switch), while gestures are of a representational nature (e.g. iconic gestures representing a trajectory).
As a first step towards natural 2D pen-based gesture recognition, we will explore the pen-based input data to be collected in the human-human experiments that are now being conducted in SLOT, with the goal to detect a limited set of salient gesture shapes. This work will be performed by WP2. Once such a set of gesture classes is distinguished, it will be examined which gesture decoders and what feature schemes are most suitable for the reliable recognition of the selected gesture classes.
In the section about human factors experiments, we introduce our plans to work on basic multi-modal experiments in the context of the graphical manipulation of iconic or geometric objects. It is expected that the techniques for constrained gesture recognition may be of use for pen-based gesture recognition in this context.

1.3.2 Towards natural 3D gesture recognition

3D gesture recognition differs from 2D as it requires different data acquisition hardware, different signal processing, segmentation and feature extraction algorithms. Whereas AHR builds on several decades of research, the first papers about automatic 3D gesture recognition appeared only after 1985. The main reason for this is that the input devices required for data capture (such as the data glove, video trackers or optotrak) still had to be invented.
Similar to natural 2D gestures, the gesture shapes that humans produce in a certain context can only be determined by thorough examination, e.g. via transcription of video material as anticipated in COMIC. Please note that the automatic recognition of natural gestures is considered as an unsolved problem. Even humans are not good at recognizing gestures as was shown in [P.Feyereisen et al., 1988], where people found it very difficult to attribute meaning to gestures based on shown videotapes, without speech or contextual clues.
However, a large number of research programs are currently being undertaken to recognize a limited number of gestures. An overview of the work in this field is presented in [Cohen, 1999]. In most cases, gesture recognition is performed in the context of virtual reality applications and control, where recognition from a selected gesture repertoire is considered as feasible.
In [Cohen, 1999], a number of architectural guidelines are specified, where the first requirement is: "Choose gestures which fit a useful environment", indicating that the research towards natural gesture recognition is still in its infancy. Furthermore, note that the design of a suitable set of gestures that are (i) easy for computers to recognize and (ii) for humans to learn, remember and reproduce is considered as particularly difficult[Jr. et al., 1999].

1.3.3 Gesture classes

As a first step towards natural gesture recognition, an inventory of the possible gesture classes is required. In [Benoit et al., 1998,McNeill, 1992,Nespoulos et al., 1986,Ruiter, 2000], various taxonomies are proposed. For the automatic interpretation of gestures, the pragmatic approach as discussed by de Ruiter [Ruiter, 2000] is adopted here:

1. Iconic gestures: depicting aspects of the accompanying speech topic. This category includes what McNeill [McNeill, 1992] calls *metaphoric* gestures, because from the perspective of gesture production it is of no relevance whether the imagery underlying the gesture is related to abstract or real entities. An example of an iconic gesture is a downward corkscrew motion of the hand while talking about a vortex.
2. Pantomimes: gestures that are imitations of functional motor activities. Example: pretending to hold a telephone while talking about a phone conversation.
3. Deictic gestures: Pointing gestures. Example: pointing somewhere and say "over there".
4. Beat gestures: Biphasic movements of the hands of finger that do not represent anything. Often clearly seen when speakers are agitated and trying to convince their audience of something.
5. Emblems: Gestures whose form-meaning relation is lexicalized. Example: the ring-shaped gesture formed by the thumb and index finger to indicate "OK".

The possibilities for automatically registering and interpreting real (i.e. naturally occurring) 3D gestures are seriously constrained by our limited understanding of how humans use and interpret gestures. As Benoit observed: "finding sets of useful gestures will probably remain an application specific development effort until gesture-based interaction will be understood in depth". However, following the gesture taxonomy presented above, a number of observations about the likelihood of successfully recognizing 3D gestures (AGR) can be made. First, the best candidates for AGR are emblems and deictic gestures. Emblems are relatively easy to recognize because they are lexicalized, meaning that one can *train* recognizers on a specified lexicon of emblem-gestures, similar to ASR where recognizers can be trained on a corpus of annotated utterances. Deictic gestures are relatively easy to interpret because of their simple semiotic properties, and their predictable relation to the concurrent speech. They can often be recognized by the

hand-shape, (the pointed finger) and a 3-dimensional vector with the speaker as origin can easily be extracted from the recorded motion data.
Far more difficult to recognize are iconic and beat gestures. Iconic gestures are by definition *not* lexicalized [McNeill, 1992], meaning that they are not governed by linguistic conventions. Usually, they are very hard to interpret (if at all) without the accompanying speech. The "iconic gesture recognition" reported in the literature [Sowa and Wachsmuth, 2001] , are usually about limited sets of predefined iconic gestures, implying that they are artificially lexicalized. In other words, they behave like (local) emblems. Finally, beat gestures are difficult to handle because nobody seems to have a clue about what their communicative function is. McNeill [McNeill, 1992] claims that they have a "meta-narrative" function, but his evidence is limited and dependent on a highly subjective interpretation of the context surrounding the utterances he collected.
It can be expected that for both COMIC applications (SLOT and ViSoft) mainly deictic gestures will be produced by human subjects. In the bath room application of ViSoft, we may also expect geometrical shapes in the form of emblems.

1.4 3D data capture and recognition

The technical annex of COMIC states that research will be conducted towards the recognition of a limited number of salient 3D pen-based gesture classes. It is expected that the analysis of the proposed human-human experiments will yield at least a workable taxonomy. Given such a set of (unconstrained, or natural) 3D gesture classes, it is intended to perform a small experiment to collect real 3D data using the optotrak 3020 system available at the NICI. This system is wall-mounted and uses three optical cameras and special hardware for capturing 3D data emitted through synchronized light-emitting active markers. The technical specifications [Inc., 1991] state a spatial accuracy of 0.15 mm and a resolution of 0.01 mm at a distance of 2.5 m between markers and the cameras. The maximum marker rate is 3500Hz, which has to be divided by the number of active markers employed. The NICI has extensively used the optotrak for examining human motor production with up to twelve markers, which still sustain a data acquisition rate of about 200Hz. Note that this is rather impressive at the given resolution.
It is yet unclear how experiments for 3D data collection will be set up. Experience has shown that users can properly perform targeted movements. How active marker technology (implying that each marker is connected by a thin wire to the hardware) restricts the unconstrained production of 3D gestures is an issue to be considered.
Several attempts towards 3D gesture recognition can be distinguished here:

- For the recognition of certain classes of deictic gestures, it will be examined whether a proper projection from 3D to 2D yields 2D-trajectories that can be recognized with the given 2D gesture recognition algorithms.

- Attempts will be made to change the existing 2D algorithms such that they are suited for the recognition of 3D trajectories. Please note that this involves a major investment in both theory, experimental evaluations and software development. For example, as was indicated above, 2D trajectories are generally normalized on the basis of holistic word-shape features. How could the baseline detection algorithm that forms the core of the normalization process be used in 3D?

- The development of new 3D pattern recognition techniques, including the evaluation of the required feature extraction algorithms. Based on a further review of the literature and the expected collection of training data, in a later stage of the COMIC project, attempts will be made to investigate this approach.

- It may become possible to describe a set of well-defined prototypical gestures in terms of elementary 3D movements. Similar to the grammar descriptions for describing character shapes, this requires some notion of velocity-based strokes in 3D. The expectation is that it will be feasible to adapt the existing velocity-based segmentation algorithms to 3D.

The OGI team has investigated several applications where 3D gestures were considered. In a review paper from Oviatt et al [Oviat et al., 2001], it is stated that 3D gesture

recognition is still beyond the state of the art. In COMIC, the joint cooperation between experts from cognitive psychology, speech and gesture recognition, computer science and artificial intelligence will hopefully provide the synergy to proceed some important steps to proceed beyond the state of the art.
Research in multimodal interaction with 3D gesture recognition is proceeding rapidly, but almost exclusively in the context of immersive applications. Much of this work is being carried out at OGI and at the MIT Media Lab. Some of those applications, such as the real estate agent REA of MIT Media Lab, may acquire a role in the field of eCommerce and eWork. Entertainment is, of course, another market segment where immersive applications will flourish. However, for the kind of applications that COMIC is focusing on combination of `simple' input and output devices, such as speech, pen and graphics the research alluded to is probably mainly of interest as far as multimodal integration is concerned [Bickmore and Cassel, 2002,Corradini and Cohen, 2002].

1.5 How to proceed beyond the state of the art?

In section 1.1.10 it is stated that larger amounts of training data and increasing technological capabilities in terms of computer power and memory lead to better trained and faster systems. This is definitely true for all areas of pattern recognition. However, a combination of new models of human cognition and motor control, improved pattern recognition techniques, and the fusion of heterogeneous, multi-modal, information provided by the input decoders is further required to lay the basis for the ambient intelligence landscapes that we envisage in COMIC.
Although ASR techniques have not basically changed much during the last decade, a number of recent developments that are being pursued in pattern recognition may speed up this research in such a way that important progress can be made. As indicated above, the frontiers in multiple classifier combination are not reached yet by far. Furthermore, we have seen a merger of techniques from various disciplines in the context of shape matching algorithms that may yield important new insights in the recognition of pen-based gestures. Also, AHGR can benefit from research from the speech recognition communities, in particular with regards to the recognition of word sequences.
Given the state-of-the-art as describe above, we conclude by enumerating a number of steps that may be followed to proceed beyond that. We do not state that we will reach all goals that are set below, but it is meant to stipulate a number of research directions that are promising enough to be followed:

- Efforts are now being made to open up HTK with the possibility to yield early probabilistic outputs that can be fused with the AGR and AHR modules, given the states in the dialog that are fed back by the dialog and action management module;
- Vice versa, research will be conducted to design AGHR modules that detect pen-input classes before the end of an input token is detected. Using knowledge from human motor control, this could also be performed for 3D gesture recognition;
- At least for AHGR, new classifiers such as support vector machines and fast hierarchical search algorithms have emerged;
- also for AHGR, the combination of online and offline data gives a solution in cases where recognition on the basis of either input is ambiguous;
- investigate the use of different feature extraction and classification techniques for (selected classes of) unconstrained gestures;
- investigate the meaning of non-linguistic 2.5D movements above the tablet;
- investigate the relation between the three input modules, with special focus on the alignment between the recognition tokens in different modalities;
- the number of basic human factors experiments in the context of the graphical manipulation of iconic or geometric objects *with a working multi-modal system* is still limited. It is the goal of workpackage 3 to develop a KUN A?R system using the agent architecture to provided by the DFKI, called *PAP* (pool architecture platform). This proof of concept system will provide an experimentation platform for several of such experiments.

1.6 Bibliography

[Bengio et al., 1994] Bengio, Y., Le Cun, Y., and Henderson, D. (1994). Globally trained handwritten word recognizer using spatial representation, convolutional neural networks, and Hidden Markov Models. In Cowan, J. D., Tesauro, G., and Alspector, J., editors, *Advances in Neural Information Processing Systems*, volume 6, pages 937-944. Morgan Kaufmann Publishers, Inc.

[Benoit et al., 1998] Benoit, C., Martin, J., Pelachaud, C., Schomaker, L., and Suhm, B. (1998). Audio-visual and multimodal speech systems. In D. Gibbon (Ed.) Handbook of Standards and Resources for Spoken Language Systems - Supplement Volume, to appear.

[Bickmore and Cassel, 2002] Bickmore, T. and Cassel, J. (2002). Phone versus face-to-face with virtual persons. In *CLASS workshop*, Copenhagen.

[Cohen, 1999] Cohen, C. (1999). A brief overview of gesture recognition. http://www.dai.ed.ac.uk/CVonline/LOCAL_COPIES/COHEN/gesture_overview.html.

[Corradini and Cohen, 2002] Corradini, A. and Cohen, P. (2002). On the relationships among speech, gestures, and object manipulation invirtual environments: Initial evidence. In *CLASS workshop*, Copenhagen.

[Duin and Tax, 2000] Duin, R. and Tax, D. (2000). Experiments with classifier combining rules. In Kittler, J. and Roli, F., editors, *Multiple Classifier Systems*, pages 16-29.

[Fonseca and Jorge, 2000] Fonseca, M. J. and Jorge, J. A. (2000). Experimental evaluation of an on-line scribble recognizer. In *Proceedings of the 11th Portuguese Conference on Pattern Recognition (PRL-RECPAD'00)*.

[Fujitsu, 2001] Fujitsu (2001). Analog resistive film input panel. website. url: http://www.fujitsu.takamisawa.com/pdf/FID550.pdf.

[Goldberg, 1989] Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, MA.

[Guyon et al., 1994] Guyon, I., Schomaker, L., Plamondon, R., and Liberman, R.and Janet, S. (1994.). Unipen project of on-line data exchange and recognizer benchmarks. In *Proceedings of the 12th International Conference on Pattern Recognition, ICPR'94*, pages 29-33, Jerusalem, Israel. IAPR-IEEE.

[Inc., 1991] Inc., N. D. (1991). Technical product description: Optotrak.

[Jaeger et al., 2000] Jaeger, S., Manke, S., and Waibel, A. (2000). Npen++: An online handwriting recognition system. In Schomaker, L. and Vuurpijl, L., editors, *Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition*, pages 249-260, Nijmegen.

[Jr. et al., 1999] Jr., A. C. L., Landay, J. A., and Rowe, L. A. (1999). Implications for a gesture design tool. In *CHI*, pages 40-47.

[Julia and Faure, 1995] Julia, L. and Faure, C. (1995). Pattern recognition and beautification for a pen based interface. In *ICDAR*, pages 58-63, Montreal.

[Klein Breteler, 2000] Klein Breteler, M. (2001). *Biophysical and cognitive determinants of human trajectory formation*. PhD thesis, University of Nijmegen.

[Lindeberg, 1998] Lindeberg, T. (1998). Edge detection and ridge detection with automatic scale selection. Technical Report ISRN KTH/NA/P-96/06-SE, CVAP, KTH, CVAP, KTH, S-100 44 Stockholm, Sweden.

[McNeill, 1992] McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago, IL.

[Mokhtarian et al., 1996] Mokhtarian, F. ., Abbasi, S., and Kittler, J. (1996). Efficient and robust retrieval by shape content through curvature scale space. In *International Workshop on Image Databases and MultiMedia Search*, pages 35-42.

[Nespoulos et al., 1986] Nespoulos, J.-L., Perron, P., and Lecours, A. (1986). *The Biological Foundations of Gestures: Motor and Semiotic Aspects*. Lawrence Erlbaum Associates.

[Oviat et al., 2001] Oviat, Cohen, Wu, Vergo, Duncan, Suhm, Besr, Holzman, Winograd, Landay, Larson, and Ferro (2001). *Designing the user interface for multimodal speech and pen-based gesture applications: state-of-the-art and future directions*, chapter 19, pages 421-456. John Carrol, Addison-Wesley.

[Parizeau et al., 2001] Parizeau, M., Lemieux, A., and Gagn, C. (2001). Character recognition experiments using unipen data. In *ICDAR*, pages 481- 485, Seattle.

[P.Feyereisen et al., 1988] P.Feyereisen, de Wiele, M. V., and Dubois., F. (1988). The meaning of gestures: What can be understood without speech? *European Bulletin of Cognitive Psychology*, 8:3(25).

[Plamondon et al., 1999] Plamondon, R., Lopresti, D., Schomaker, L., and Srihari, R. (1999). On-line handwriting recognition. *Wiley Encyclopedia of Electrical & Electronics Engineering*, pages 123-146.

[Ratzlaff, 2001] Ratzlaff, E. (2001). A scanning n-tuple classifier for online recognition of handwritten digits. In *ICDAR*, pages 18-22, Seattle.

[Rubine, 1991] Rubine, D. (1991). Specifying gestures by example. *ACM Journal on Computer Graphics*, 25 (4):329-337.

[Ruiter, 2000] Ruiter, J. d. (2000). The production of gesture and speech. In McNeill, D., editor, *Language and gesture*.

[Schomaker, 1993] Schomaker, L. (1993). Using stroke- or character-based self-organizing maps in the recognition of on-line, connected cursive script. *Pattern Recognition*, 26(3):443-450.

[Schomaker and Teulings, 1990] Schomaker, L. and Teulings, H.-L. (1990). A handwriting recognition system based on the properties and architectures of the human motor system. In *Proceedings of the International Workshop on Frontiers in Handwriting Recognition (IWFHR)*, pages 195-211, Montreal: CENPARMI Concordia.

[Schomaker et al., 1999] Schomaker, L., Vuurpijl, L., and de Leau, E. (1999). New use for the pen: outline-based image queries. In *Fifth International Conference on Document Analysis and Recognition*, pages 293-296. IEEE.

[Sowa and Wachsmuth, 2001] Sowa, T. and Wachsmuth, I. (2001). Interpretation of shape-related iconic gestures in virtual environments. In *Gesture and Sign Language in Human-Computer Interaction*, International Gesture Workshop, GW2001. Springer-Verlag.

[Srihari et al., 2001] Srihari, S., Cha, S.-H., and Lee, S. (2001). Establishing handwriting individuality using pattern recognition techniques. In *ICDAR*, pages 1195-1204, Seattle.

[Steinherz et al., 2002] Steinherz, T., Rivlin, E., and Intrator, N. (2002). Off-line cursive script word recognition: A survey. *International Journal of Document Analysis and Recognition*, 2(2). (to appear).

[Veltkamp and Hagedoorn, 2000] Veltkamp, R. and Hagedoorn, M. (2000). Shape matching: Similarity measures and algorithms. Technical report, Utrecht University. Technical Report UU-CS-.

[Vinciarelli, 2002] Vinciarelli, A. (2002). A survey on off-line cursive script recognition. *Pattern Recognition*, 35(7):1433-1446.

[Vuurpijl, 2002] Vuurpijl, L. (2002). Annotated website with links to multi-modal speech and pen-based gesture recognition. http://hwr.nici.kun.nl/~vuurpijl/comic.

[Vuurpijl and Schomaker, 1997] Vuurpijl, L. and Schomaker, L. (1997). Finding structure in diversity: A hierarchical clustering method for the categorization of allographs in handwriting. In *ICDAR*, pages 387-393. IEEE.

[Vuurpijl and Schomaker, 1998] Vuurpijl, L. and Schomaker, L. (1998). Multiple-agent architectures for the classification of handwritten text. In *IWFHR6, International Workshop on Frontiers of Handwriting Recognition*, pages 335-346.

[Vuurpijl and Schomaker, 2000] Vuurpijl, L. and Schomaker, L. (2000). Two-stage character classification: a combined approach of clustering and support vector classifiers. In Schomaker, L. and Vuurpijl, L., editors, *Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition*, pages 423-432, Nijmegen. iUF/NICI, International Unipen Foundation. (paperback, 620 pages, illustrated).

[Vuurpijl et al., 2002a] Vuurpijl, L., Schomaker, L., and van Erp, M. (2002a). Architectures for detecting and solving conflicts: two-stage classification and support vector classifiers. *International Journal of Document Analysis and Recognition*. Submitted to the special issue on multiple classifiers systems for document analysis applications.

[Vuurpijl et al., 2002b] Vuurpijl, L., Schomaker, L., and vd Broek, E. (2002b). Vind(x): using the user through cooperative annotation. In *IWFHR8, International Workshop on Frontiers of Handwriting Recognition*, pages 221-226.

[Wang et al., 2000] Wang, F., vuurpijl, L., and Schomaker, L. (2000). Support vector machines for the classification of western handwritten capitals. In Schomaker, L. and Vuurpijl, L., editors, *Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition*, pages 167-176, Nijmegen. iUF/NICI, International Unipen Foundation. (paperback, 620 pages, illustrated).

[Wu et al., 1999] Wu, L., Oviatt, S. L., and Cohen, P. R. (1999). Multimodal integration - a statistical view. *IEEE Transactions on Multimedia*, 1(4):334-341.

2 State of the Art in Natural Language Understanding and Multimodal Fusion

Ralf Engel, Norbert Pfleger DFKI GmbH

This chapter describes the state of the art in natural language understanding (NLU) and multimodal fusion for spoken dialogue systems. The aim is to give an broad overview of the currently used approaches.

2.1 State of the art in natural language understanding (NLU)

2.1.1 Overview

The task of the natural language understanding (NLU) component within a dialogue system is to extract the semantic content of the spoken utterance. The input of the NLU component is the output of the speech recogniser (ASR), either the best hypothesis, an n-best list of scored hypotheses or a word lattice. Some recognisers also add prosodic information including boundary information. The output is usually sent to the dialogue manager (DM) and is represented using frames, typed feature structures or terms in some logic calculus. Many NLU components extract not just one representation but several scored alternatives due to different hypotheses send by the speech recogniser and ambiguous words / expressions within one hypothesis.
Since techniques developed originally for written text could be also adapted for NLU in spoken dialogue systems, not only approaches actually used in dialogue systems are presented in this report but rather a broader overview of NLU approaches is given. Applications using an NLU component can be divided into the following groups:

- Information extraction (IE) including applications for automatic e-mail and call type classification and text summarization.
- Dialogue systems including information systems, e.g., for television program, tourist information and planning systems, e.g., military simulations.
- Machine Translation (MT).
- Chatterbots providing some kind of natural conversation to entertain people, e.g., for advertising purposes.

For these tasks various approaches for NLU were developed. The most important properties in which these approaches differ are:

Coverage / Universality

The spectrum reaches from approaches designed to fulfil highly domain specific tasks, like proper name recognition to approaches claiming to be able to process all natural language input in a satisfactory manner.
The approaches differ also in how well different languages and their idiosyncrasies are supported, e.g., free constituent order in German.

Robustness

The ability to handle unexpected input. Sources for unexpected input include:

- Recognition errors in ASR, speaker independent ASR systems still have an word error rate (WER) of about 20%.

- The user can use words that are not in the vocabulary of the ASR which leads to incorrect or incomplete (provided the ASR is able to detect unknown words) recognition
- Semantically unexpected input, e.g., the user combines concepts in an unexpected manner or overestimates the capabilities of the system.
- Syntactically incorrect utterances / text including hesitations, corrections etc., this is especially a problem when dealing with spontaneous spoken input.
- Syntactic complexity of the input exceeds the capabilities of the system.

Kind of output

Usually the output format consists either of frames, typed feature structures which differ from frames only in allowing structure sharing or of terms written in some logic calculus, e.g., DRT. The approaches also differ in which degree the structure of the output is restricted or can be varied by the knowledge bases, i.e., to what degree the structure of the output is build in or defineable in the knowledge bases.
Some approaches also support a compact representation for multiple / ambigious output. This can reduce the size of the output dramatically since in many cases local alternatives are independent and so the alternatives are multiplied if no compact representation exists.

Ease of creation and extension of knowledge bases

Knowledge bases usually include a lexicon, grammar rules and semantic knowledge. Different approaches can have different curves how fast the knowledge bases typically grow for a given point in the development cycle. E.g., some approaches support a quick start but extensions can quickly get very complex or the start is more time consuming but then extensions are less complex. Some approaches also support learning from annotated corpora or let the users extend the knowledge bases themselves, e.g., by offering short clarification dialogues.

Ability to cooperate with ASR systems

Some approaches can be tightly coupled with ASRs to decrease the word error rate by integrating the NLU in the internal search strategies. Others are optimized for processing n-best lists or word lattices and use syntactic / semantic knowledge to score the different alternatives. Some NLU approaches also allow the offline creation of language models for the ASR by exploiting their knowledge bases.

Performance

In interactive systems like dialogue systems the performance is to some extend crucial. If the approach is fast enough and the ASR produces n-best lists or word lattices it is also possible to process several alternatives and choose the most appropriate one using a scoring function.
For performance measurement different issues have to be considered: The time until the first solution is generated, the time until all solutions are generated and if ``more probable'' solutions are found first.
Unfortunately not all approaches are described detailed enough in the corresponding literature and manuals to give satisfactory answers to all items.

2.1.2 Approaches

Research in the field of NLU is presented at various conferences, e.g., the meeting of the Association for Computational Linguistics (ACL), the International Conference on Speech and Language Processing (ICSLP), European Conference on Speech Communication and Technology (Eurospeech), the conference on Applied Natural Language Processing (ANLP), the International Conference on Computational Linguistics (COLING), the Human Language Technology Conference (HLT), the International Workshop on Parsing Technologies (IWPT) and Empirical Methods in Natural Language Processing (EMNLP). Also a large number of workshops are organized on various topics, e.g., robust parsing.

Since there is a long ongoing research over the last 30 years, presentations of radically new approaches become quite rare. In most cases variations or new combinations of established approaches are presented.
For the task of NLU two principal approaches are available, concept spotting and grammar-based parsing.

Concept spotting

Concept spotting means that only portions instead of the whole text / utterance are analysed. These portions are used for semantic slot-filling of predefined frames (in IE also called templates). This approach is also known as case based parsing.
To extract portions of the text / utterance, usually manually written regular expressions or finite state transducers (FSTs) [Roche and Schabes, 1997] are used in combination with a semantically annotated lexicon.
Concept spotting provides robust and fast processing, but is not intended to get the exact full meaning of the text / utterance and is highly domain dependent. Since only portions of the utterance are extracted, this approach is rather unsuitable to be tightly coupled with ASR. Concept spotting is used within IE systems, dialogue systems and chatterbots.
Currently a lot of research is done to avoid the time consuming manual creation of patterns / FSTs. One approach is using a large annotated corpus and learn the patterns / FSTs automatically using, e.g., hidden Markov models [Minker et al., 1996],[Schwarz et al., 1997], neural nets [Wermter and Weber, 1994] or relational learning methods [Califf, 1999]. A drawback is that such a large annotated corpus must already exist or has to be build from scratch. Another approach is the semi- automatic generation of patterns using term extraction on a non-annotated corpus in conjunction with an ontology [Maedche et al., 2002].

Implementations

IE systems currently under development are CICERO (Language Computer Corporation) using patterns and SMES [Neumann et al., 1997] using FSTs.
A very popular FST-parser used in dialogue systems is the Phoenix parser [Ward, 1991] developed at Carnegie Mellon University (CMU) and used in many projects including the CU Communicator [Pellom et al., 2001] and Carnegie Mellon Communicator [Constantinides and Rudnicky, 1999]. FSTs are used for detecting semantic entities which are inserted in frames. Although the parser is old, it is still used and further enhanced.
Other parsers based on FSTs are the GEM parser [Miller and Stallard, 1996] used by BBN for the Talk'n'Travel system [Stallard, 2000] and the fall-back parser WORDSPOT [Burianek, 2000] used in the Galaxy Communicator [Polifroni and Seneff, 2000].

Grammar-based parsing

The whole text / utterance is analysed using a syntactic and / or semantic grammar. The produced output is usually either frame-based or a term expressed in a logic calculus, e.g., DRT. This approach is usually less robust and slower than concept spotting (although much research is done to address these issues), but provides a more exact representation of the text / utterance. Dependent on the implementation the systems may be domain dependent or not. Grammar-based parsing is used for IE (but not so common as concept spotting), dialogue systems and MT.
For the use within dialogue systems, context free grammars (CFGs) extended with feature structures are still regarded as sufficient. There are one-stage and two-stage approaches. In the one-stage approach, a semantic CFG or an offline interleaved syntactic and semantic CFG is used. In the two-stage approach, first a CFG for syntactic analysis is used and in a second step, the semantic content is extracted from the parse tree using simple rules. One-stage semantic grammars have the advantage of limiting ambiguity using their semantic knowledge and thereby enable faster parsing. The grammar is processed either directly by a chart based or depth-first top-down parser or the grammar is compiled offline into FSTs and the FSTs are used for the actual parsing.

This allows an easy direct integration into the A*-search in the ASR although the results are not very promising, e.g. [Mangu et al., 1999].
As already mentioned, two drawbacks of CFGs have to be addressed: speed and robustness. The performance can be increased by using only a subset of CFG which is regarded to be sufficient for the task of NLU and allow the use of faster parsing algorithms, like GLR [Tomita, 1987], or by annotating the rules with statistical information extracted from large, not necessarily annotated corpora to prune unlikely paths, e.g., [Potamianos and Kuo, 2000] and [Kaiser et al., 1999]. The statistical information can also be used when the parser is integrated in the ASR.
Robustness can be addressed by enhancing the grammar with ``garbage'' rules that skip unsuitable words or by modifying the parser to be able to deliver partial analyses in the case full parses are not found, e.g., GLR* [Lavie, 1996] or by asking clarifying questions to the user [Rosé, 1997].
Like in concept spotting approaches, research is done in the field of data-oriented rule learning, e.g., at the University of Amsterdam [Bod, 1998] and the University of Karlsruhe [Buo and Waibel, 1996]. Instead of using an annotated corpus there is also ongoing research learning new knowledge with the help of the user [Gavaldà, 2000a].
To provide a larger coverage of natural language, constraint-based grammar formalisms, like HPSG [Pollard and Sag, 1994], LFG [Dalrymple, 1999] and TAG [Joshi and Schabes, 1997] are used. The central formalism used is unification of feature-structures. To be used in spoken dialogue systems their handling is still too complex and performance and robustness are not sufficient. The OVIS-2 project at the University of Groningen tries to combine a complex grammar approach with robustness and efficiency using a Definite Clause Grammar (DCG) [Noord et al., 1999].

Implementations

Examples for CFG-parsers developed for dialogue systems and used in currently running or recently finished projects are:

The TRIPS-parser [Allen, 1994]

The parser used in the TRIPS system [Allen et al., 2001] was developed at the University of Rochester. It uses a syntactic and semantic grammar which is combined to be used in an one stage parser. The parser is a bottom-up chart parser which can produce partial parse results for robustness and can work on word lattices.

TINA [Seneff, 1992]

The parser was developed at MIT and is used there and in other sites for many projects, e.g., Jupiter [Zue et al., 2000] and SpeechBuilder [Glass and Weinstein, 2001]. The grammar is a probabilistic semantically tagged CFG, the parser is a top-down chart processor. Recent research is on integrating the parser in an ASR [Lau and Seneff, 1998] and on considering confidence values delivered by the ASR in the word lattice for a more robust dialogue system [Hazen et al., 2000].

PROFER [Kaiser et al., 1999]

The parser is under development at the OGI. The grammar is a probabilistic CFG which is compiled into FSTs to be tightly coupled with an ASR.

SOUP [Gavaldà, 2000b]

The parser was developed at Carnegie Mellon University and is used, e.g., in LingWear [Fügen et al., 2001]. It is a stochastic, chart-based top-down parser. The CFGs are compiled into recursive FSTs. Robustness is achieved by allowing skipping of input words at any position and producing ranked interpretations that may consists of multiple parse trees.

2.1.3 How do we proceed beyond the state of the art?

Due to robustness and performance issues and the lower complexity of spoken utterances, the current state of the art in NLU for spoken dialogue systems is still characterised by rather simple but effective approaches instead of complex unification

based approaches like HPSG. Within the SmartKom project we developed a state of the art template based NLU module.
Beside other improvements on the module the following two issues will be addressed in COMIC that are beyond the current state of the art of NLU modules in spoken dialogue systems.
First, coordination of phrases which contain ellipses, e.g., ``Place this toilet there and the basin here,'' will be supported. Such constructions are quite common in spoken dialogues.
Second, it should be possible to extend the knowledge bases to cover more variant phrasings just using example utterances instead of modifying the knowledge bases directly. The past has shown that this is one of the most time consuming tasks and by restricting ourselves to variant phrasings, we hope to reduce the complexity of such an automated process enough to get it working.

2.2 State of the art in multimodal fusion

2.2.1 Motivation

Multimodal dialogue systems enable the user to interact with the system using different modalities at the same time. This interaction is usually based on speech combined with mouse/pen gestures, 2D/3D hand gestures, or haptic devices. Commonly, complex gestures like arrows, underlining, drawing symbols and digits are supported besides simple pointing. Few systems also support handwriting. Some systems deal also with the integration of other modalities, like facial expression recognition, e.g., for mood detection and lip reading supporting the ASR.
Using multimodal interfaces, the user is able to combine the advantages of the individual input modalities or switch the modalities depending on the environment. Speech offers the ability to specify complex tasks with relative ease and permits the user to interact with their ``hands off.'' Gestures offer simple mentioning of objects already shown on the screen, triggering of simple actions either by drawing symbols or pointing on icons on the screen and conveying spatial information of points, lines and areas. Gestures are preferred over speech in noisy or quiet environments.
As shown by Oviatt [Oviatt, 1999] multimodal input can also be used for resolving ambiguous input in one modality and for supporting the detection and correction of recognition errors. In this study the processing results of over 2,000 multimodal utterances performed by both native and accented speakers were logged and analysed. The results of this study confirm that multimodal systems can support mutual disambiguation significantly. Also users tend to use a simplified language in combination with gestures resulting in more robust natural language processing. Besides this, users tend to prefer the less erroneous modality for a specific task.

2.2.2 Exploring the Integration and Synchronization of Input Modalities

Oviat et al. [Oviatt et al., 1997] examined mulitmodal interaction where people speak and write to a simulated dynamic map system. They analysed data collected from 72 map interaction tasks performed by eighteen native English speakers. The resulting multimodal corpus was coded for the following dependent measures:

- *User Preference* The percentage of users preferring to interact unimodally or multimodally during map tasks was summarized.

- *Task Actions* The user commands found in the corpus were classified into types of action commands out of a set of 14 types.

- *Linguistic Content* They analysed the order of semantic constituents such as subject, verb, object, and locatives to determine whether the input order of multimodal constructions conforms with the canonical S-V-O ordering expected for English.

- *Multimodal Integration Patterns* Finally, the constructions performed by the subjects were classified into a set of integration patterns.

In the following subsection we will give a brief summarization of the results of the study of Oviatt et al.

Results

User Preference

Subjects showed a strong preference to interact multimodally during map tasks (100% of them used at least once a combination of spoken and pen input during one task). 19% of the constructions were expressed multimodally, 17.5% unimodally through writing, and 63.5% using only speech.

Task Action Analysis

Spatial location commands that require a spatial location description are the user commands that were most likely expressed multimodally (86% of the multimodal constructions). Commands involving the selection of a specific object on the screen accounted for 11% of the users' multimodal constructions. However, it turned out that commands that involve the selection of an in-view object were more likely expressed unimodally, because the object was already in focus from the previous dialogue context or the visual context.
Other types of user commands were rarely expressed multimodally (only 3% of the multimodal constructions).

Linguistic Content

98% of the multimodal construction conformed with the standard S-V-O order typical of English (compared to 98% of the unimodal spoken constructions). The prominent difference between spoken and multimodal constructions was the typical position of locative constituents. 96% of the spoken utterances showed locatives at sentence-final position, whereas in 95% of the multimodal construction locatives (using the pen) were first expressed and followed by spoken S-V-O constituents.
Only 41% of the multimodal utterances did contain a spoken deictic term.

Multimodal Integration Patterns

86% of the multimodal constructions showed a draw and speak pattern, where in 42% a simultaneous integration of drawing and writing, in 32% a sequential input, and in 12% a compound pattern took place. The remaining 14% of multimodal constructions showed a point and speak pattern.
The synchronization patterns for the simultaneous integration of the draw and speak condition showed in 57% a precedence of writing and only in 14% a precedence of speech.
Sequential integration patterns showed a temporal precedence of written input in 99% of the cases. The lag between the end of the writing and the start of speech averaged 1.4 seconds, with an maximum lag of 4 seconds.
The integration of deictic terms showed in 43% a sequential integration pattern with an average lag of 1.1 seconds (a maximum lag of 3.0 seconds in 97% of the time).

2.2.3 Requirements for multimodal fusion

Considering the above mentioned aspects of multimodal interaction, we can identify the following two major requirements for multimodal fusion:

- The integration patterns mentioned above have to receive attention during the synchronization of the input modalities.
- The resulting structures have to be semantically meaningful, especially unintended representations have to be ruled out (disambiguation).

2.2.4 Approaches

In the following subsection we consider the two major architectural approaches: early fusion and late fusion.

Early fusion at feature level

The recognition process in one modality influences the course of recognition in the other. To be useful, the input modalities have to be closely coupled and synchronized, like speech and lip movement. Early fusion is not well suited for modalities that differ substantially in the information content or time scale characteristics of their features, like speech and pen gestures. Additionally there is an integration overhead since the recognizers have to be modified to work together and a multimodal training corpus must exit. Examples of representative research include [Bregler et al., 1993], [Vo et al., 1995] and [Pavlovic and Huang, 1998]. Since in COMIC only speech and gestures are involved, early fusion is not further discussed in this report, but maybe gestures (especially drawn numbers and handwriting) can be used in a limited fashion to restrict the language model of the ASR.

Late fusion at semantic level

The second architectural approach is based on independent recognizers where their respective output is integrated afterwards. Usually the output of the different modalities is represented in a frame based or typed feature structure and has to be combined, resulting also in a frame based or typed feature structure. The combination process has to address the following aspects:

- The combination has to be semantically meaningful.
- The utterances and gestures have to be grouped in a time-sensitive fashion, e.g., if a gesture occurs between two spoken utterances, the gesture could be interpreted with the preceding utterance, the following utterance, or by itself.
- The combination within one group has also to be time sensitive, e.g., when the user specifies a route on a map by pointing to several locations combined with spoken elaborations. Therefore, meaning fragments need to include time stamps.
- Alternative solutions have to be scored and the best joint interpretation has to be chosen. The scoring function must be aware of recognition errors in the modalities and has to compensate these.

In order to fuse information derived from multiple modalities, various research groups including [Vo and Wood, 1996], [Cheyer and Julia, 1995], [Pavlovic and Huang, 1998] and [Shaik et al., 1997] have independently converged on a strategy of recursive matching and merging of attribute/value structures, although details of the algorithms differ. They are similar to unification of typed feature structures, which is directly used by [Johnston et al., 1997].
There is also ongoing research how to integrate statistical processing techniques [Wu et al., 1999]. A relatively new approach is to integrate multimodal fusion in the NLU, e.g., M. Johnston presents an approach using modified FSTs that can parse spoken input and gestures at the same time [Johnston and Bangalore, 2000].

2.2.5 Implementations

In the following we give a short overview over some implementations of multimodal dialogue systems:

QuickSet (OGI) [Cohen et al., 1997]

Application: Creating and positioning entities and supplying their behavior in distributed, interactive simulations (DIS) training environments for battlefield simulations. QuickSet runs on a hand-help PC. Typed feature structure unification is used for multimodal fusion.

Virtual Reality Aircraft Maintenance Training Prototype (Boeing) [Duncan et al., 1999]

The system is intended for use in assessing the maintainability of new aircraft designs and training mechanics in maintenance procedures using virtual reality. The system uses a Cyberglove gesture input device and a 3D-gesture recognizer. Speech and gesture input is combined by filling time-stamped gestural frames into time-stamped speech event frames.

LingWaer [Fügen et al., 2001]

LingWear is a mobile tourist information system. The user can communicate with LingWear either by means of spontaneous speech queries or via touch screen.

Portable Voice Assistant (BBN) [Bers et al., 1998]

Speech and gesture (including handwriting) on a hand-held device realized as Java-applet in a web browser. The first prototype application is an on-line vehicle repair manual and parts ordering systems. Speech and pen input events are integrated to a frame-based description of the user's request.

Match [Johnston et al., 2002]

Match is a mobile multimodal speech-pen interface for retrieving restaurant and subway information for New York City. Based on finite-state methods, the system enables the user to interact using pen, speech or a combination of them. The architecture of MATCH consists of a set of agents which communicate through a java-based facilitator MCUBE.

SmartKom [Wahlster et al., 2001]

SmartKom (see `www.smartkom.org`) is a multimodal dialogue system combining speech, gesture and facial expression input and output. The user can operate this system by using a combination of spoken language, gestures and facial expressions. On the output side an animated life-like character (called *Smartakus*) is displayed on the screen who interacts with the user by combining graphical output with gestures and speech. The SmartKom dialogue system is currently under development at the DFKI in cooperation with several academic and industrial partners.

2.2.6 How do we proceed beyond the state of the art?

We plan to build on our work of multimodal fusion within the SmartKom system. In a first step we extend our approach to the pen-based interface of Comic. In the following steps we will concentrate on the different integration patterns identified by Oviatt et al. and representational and computational aspects of multimodal fusion.

2.3 Bibliography

[Allen, 1994] Allen, J. (1994). *Natural Language Understanding*. Benjamin/Cummings Publishing Co, Menlo Park, CA, 2nd ed. edition.

[Allen et al., 2001] Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L., and Stent, A. (2001). Towards conversational human-computer interaction. *AI Magazine*.

[Bers et al., 1998] Bers, J., Miller, S., and Makhoul, J. (1998). Designing conversational interfaces with multimodal interaction. In *DAPRA Workshop on Broadcast News Understanding Systems*, pages 319-321.

[Bod, 1998] Bod, R. (1998). Spoken dialogue interpretation with the DOP model. In *Proceedings of COLING-ACL-98*, Montreal, Canada.

[Bregler et al., 1993] Bregler, C., Manke, S., Hild, H., and Waibel, A. (1993). Improving connected letter recognition by lipreading. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP-93)*, pages 557-560.

[Buo and Waibel, 1996] Buo, F. D. and Waibel, A. (1996). Feaspar - a feature structure parser learning to parse spoken language. In *Proceedings of 16th International Conference on Computational Linguistics (COLING-96)*, Copenhagen, Denmark.

[Burianek, 2000] Burianek, T. (2000). Building a speech understanding system using word spotting techniques. Master's thesis, MIT Department of Electrical Engineering and Computer Science.

[Califf, 1999] Califf, M. E. (1999). Relational learning of pattern-match rules for information extraction. In *Proceedings of 16th National Conference on Artifical Intelligence (AAAI-99)*.

[Cheyer and Julia, 1995] Cheyer, A. and Julia, L. (1995). Multimodal maps: An agent-based approach. In *International Conference on Cooperative Multimodal Communication (CMC-95)*, pages 63-69, Tilburg, The Netherlands.

[Cohen et al., 1997] Cohen, P., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., and Clow, J. (1997). Quickset: Multimodal interaction for distributed applications. In *Proceedings of the 5th ACM International Multimedia Conference*, pages 31-40.

[Constantinides and Rudnicky, 1999] Constantinides, P. C. and Rudnicky, A. I. (1999). Dialogue analysis in the Carnegie Mellon Communicator. In *Proceedings of 6th European Conference on Speech Communication and Technology (Eurospeech-99)*, Budapest, Hungary.

[Dalrymple, 1999] Dalrymple, M. (1999). Lexical-functional grammar. In Wilson, R. and Keil, F., editors, *MIT Encyclopedia of the Cognitive Sciences*. The MIT Press.

[Duncan et al., 1999] Duncan, L., Brown, W., Esposito, C., Holmback, H., and Xue, P. (1999). Enhancing virtual maintenance environments with speech understanding. *Boeing M&CT TechNet*.

[Fügen et al., 2001] Fügen, C., Westphal, M., Schneider, M., Schultz, T., and Waibel, A. (2001). Lingwear: A mobile tourist information system. In *Proceedings of Human Language Technology Conference (HLT-2001)*, San Diego, CA.

[Gavaldà, 2000a] Gavaldà, M. (2000a). Epiphenomenal grammar acquisition with GSG. In *Proceedings of the Workshop on Conversational Systems of the 6th Conference on Applied Natural Language Processing and the 1st Conference of the North American Chapter of the Association for Computational Linguistics (ANLP/NAACL-2000)*, Seattle.

[Gavaldà, 2000b] Gavaldà, M. (2000b). SOUP: A parser for real-world spontaneous speech. In *Proceedings of 6th International Workshop on Parsing Technologies (IWPT-2000)*, Trento, Italy.

[Glass and Weinstein, 2001] Glass, J. and Weinstein, E. (2001). Speechbuilder: Facilitating spoken dialogue system development. In *Proceedings of 7th European Conference on Speech Communication and Technology (Eurospeech-2001)*, Aalborg, Denmark.

[Hazen et al., 2000] Hazen, T. J., Burianek, T., Polifroni, J., and Seneff, S. (2000). Recognition confidence scoring for use in speech understanding systems. In *Proceedings of ISCA ASR2000 Tutorial and Research Workshop*, Paris.

[Johnston and Bangalore, 2000] Johnston, M. and Bangalore, S. (2000). Finite-state multimodal parsing and understanding. In *Proceedings of 18th International Conference on Computational Linguistics (COLING-2000)*, Saarbrücken, Germany.

[Johnston et al., 2002] Johnston, M., Bangalore, S., Stent, A., Vasireddy, G., and Ehlen, P. (2002). Multimodal language processing for mobile information access. In *Proceedings of 7th International Conference on Spoken Language Processing (ICSLP-2002)*, Denver, Colorado.

[Johnston et al., 1997] Johnston, M., P.R., C., McGee, D., Oviatt, S., and Pittman, J.A. amd Smith, I. (1997). Unification-based multimodal integration. In *Proceedings of 35th Annual Meeting of the Association for Computational Linguistics, (ACL-97)*, pages 281-288, NewYork.

[Joshi and Schabes, 1997] Joshi, A. and Schabes, Y. (1997). *Handbook of Formal Languages*, volume 3, chapter Tree-Adjoining Grammars, pages 69-124. Springer, Berlin, New York.

[Kaiser et al., 1999] Kaiser, E. C., Johnston, M., and Heeman, P. A. (1999). PROFER: Predictive, robust finite-state parsing for spoken language. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP-99)*, volume 2, pages 629-632, Phoenix, Arizona.

[Lau and Seneff, 1998] Lau, R. and Seneff, S. (1998). A unified framework for sublexical and linguistic modelling supporting flexible vocabulary speech understanding. In *Proceedings of 5th International Conference on Spoken Language Processing (ICSLP-98)*, Sydney, Australia.

[Lavie, 1996] Lavie, A. (1996). GLR*: A robust parser for spontaneously spoken language. In *Proceedings of ESSLLI-96 Workshop on Robust Parsing*.

[Maedche et al., 2002] Maedche, A., Neumann, G., and Staab, S. (2002). Bootstrapping an ontology-based information extraction system. In Szczepaniak, P., Segovia, J., Kacprzyk, J., and Zadeh, L., editors, *Intelligent Exploration of the Web*. Physica-Verlag, Heidelberg, Germany.

[Mangu et al., 1999] Mangu, L., Brill, E., and Stolcke, A. (1999). Finding consensus among words: Lattice-based word error minimization. In *Proceedings of 6th European Conference on Speech Communication and Technology (Eurospeech-99)*, Budapest, Hungary.

[Miller and Stallard, 1996] Miller, S. and Stallard, D. (1996). A fully statistical approach to natural language interfaces. In *Proceedings of 34th Annual Meeting of the Association for Computational Linguistics (ACL-96)*, Santa Cruz, CA.

[Minker et al., 1996] Minker, W., Bennacef, S., and Gauvin, J. (1996). A stochastic case frame approach for natural language understanding. In *Proceedings of 4th International Conference on Spoken Language Processing (ICSLP-96)*, pages 1013-1016, Philadephia.

[Neumann et al., 1997] Neumann, G., Backofen, R., Baur, J., Becker, M., and Braun, C. (1997). An information extraction core system for real world german text processing. In *Proceedings of 5th Conference on Applied Natural Language Processing (ANLP-97)*, pages 209-216, Washington, D.C.

[Noord et al., 1999] Noord, G. v., Bouma, G., Koeling, R., and Nederhof, M.-J. (1999). Robust grammatical analysis for spoken dialogue systems. *Journal of Natural Language Engineering*, 5(1):45-93.

[Oviatt, 1999] Oviatt, S. L. (1999). Mutual disambiguation of recognition errors in a multimodal architecture. In *CHI*, pages 576-583.

[Oviatt et al., 1997] Oviatt, S. L., DeAngeli, A., and Kuhn, K. (1997). Integration and synchronization of input modes during multimodal human-computer interaction. In *CHI*, pages 415-422.

[Pavlovic and Huang, 1998] Pavlovic, V. and Huang, T. (1998). Multimodal prediction and classification on audio-visual features. In *AAAI'98 Workshop on Representations for Multi-modal Human-Computer Interaction*, pages 55-59.

[Pellom et al., 2001] Pellom, B., Ward, W., Hansen, J., Hacioglu, K., Zhang, J., Yu, X., and Pradhan, S. (2001). University of Colorado dialog systems for travel and navigation. In *Proceedings of Human Language Technology Conference (HLT-2001)*, San Diego, CA.

[Polifroni and Seneff, 2000] Polifroni, J. and Seneff, S. (2000). GALAXY-II as an architecture for spoken dialogue evaluation. In *Proceedings of International Conference on Language Resources and Evaluation*, Athens, Greece.

[Pollard and Sag, 1994] Pollard, C. and Sag, Ivan, A. (1994). *Head-Driven Phrase Structure Grammar*. The University of Chicago Press.

[Potamianos and Kuo, 2000] Potamianos, A. and Kuo, H.-K. (2000). Statistical recursive finite state machine parsing for speech understanding. In *Proceedings of 6th International Conference on Spoken Language Processing (ICSLP-2000)*, Beijing, China.

[Roche and Schabes, 1997] Roche, E. and Schabes, Y., editors (1997). *Finite State Language Processing*. The MIT Press.

[Rosé, 1997] Rosé, C. P. (1997). *Robust Interactive Dialogue Interpretation*. PhD thesis, Language Technologies Institute, School of Computer Science, Carnegie Mellon University.

[Schwarz et al., 1997] Schwarz, R., Miller, S., Stallard, D., and Makhoul, J. (1997). Hidden understanding models for statistical sentence understanding. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP-97)*, pages 1479-1482, Munich, Germany.

[Seneff, 1992] Seneff, S. (1992). TINA: A natural language system for spoken language application. *Computational Linguistics*, 18(1).

[Shaik et al., 1997] Shaik, A., Juth, S., Medl, A., Marsic, I., Kulikowski, C., and Flanagan, J. (1997). An architecture for multimodal information fusion. In *Proceedings of the Worshop on Perceptual User Interfaces (PUI-97)*, pages 91-93, Banaf, Canada.

[Stallard, 2000] Stallard, D. (2000). Talk'n'Travel: A conversational system for air travel processing. In *Proceedings of 6th Conference on Applied Natural Language Processing (ANLP-2000)*, Seatle, Washington.

[Tomita, 1987] Tomita, M. (1987). An efficient augmented context-free parsing algorithm. *Computational Linguistics*, 13:31-46.

[Vo et al., 1995] Vo, M., Hoghton, R., Yang, R., Meier, U., Waibel, A., and Duchnowski (1995). Multimodal learning interfaces. In *Proceedings of the DARPA Spoken Language Technology Workshop*.

[Vo and Wood, 1996] Vo, M. and Wood, C. (1996). Building an application framework for speech and pen input integration in multimodal learning interfaces. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP-96)*, pages 3545-3548, Atlanta, Georgia.

[Wahlster et al., 2001] Wahlster, W., Reithinger, N., and Blocher, A. (2001). Smartkom: Multimodal communication with a life-like character. In *Proceedings of 7th European Conference on Speech Communication and Technology (Eurospeech-2001)*, Aalborg, Denmark.

[Ward, 1991] Ward, W. (1991). Understanding spontaneous speech: the Phoenix system. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP-91)*.

[Wermter and Weber, 1994] Wermter, S. and Weber, V. (1994). Learning fault-tolerant speech parsing with screen. In *Proceedings of 12th National Conference on Artifical Intelligence*, Seattle.

[Wu et al., 1999] Wu, L., Oviatt, S. L., and Cohen, P. R. (1999). Multimodal integration - a statistical view. *IEEE Transactions on Multimedia*, 1(4):334-341.

[Zue et al., 2000] Zue, V., Seneff, S., Glass, J., Joseph, P., Pao, C., J., H. T., and Hetherington, L. (2000). JUPITER: A telephone-based conversational interface for weather information. In *IEEE Transactions on Speech and Audio Processing*, pages 100-112.